JCTC Journal of Chemical Theory and Computation

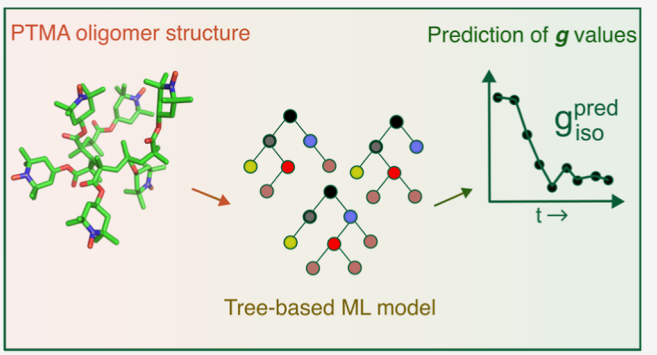
pubs.acs.org/JCTC Article

Machine Learning Isotropic *g* Values of Radical Polymers

Davis Thomas Daniel,* Souvik Mitra, Rüdiger-A. Eichel, Diddo Diddens, and Josef Granwehr

Cite This: *J. Chem. Theory Comput.* 2024, 20, 2592−2604 Read Online

ABSTRACT: Methods for electronic structure computations, such as density functional theory (DFT), are routinely used for the calculation of spectroscopic parameters to establish and validate structure−parameter correlations. DFT calculations, however, are computationally expensive for large systems such as polymers. This work explores the machine learning (ML) of isotropic *g* values, g_{iso}, obtained from electron paramagnetic resonance (EPR) experiments of an organic radical polymer. An ML model based on regression trees is trained on DFT-calculated *g* values of poly(2,2,6,6-tetramethylpiperidinyloxy-4-yl methacrylate) (PTMA) polymer structures extracted from different time frames of a molecular dynamics trajectory. The DFT-derived *g* values, g_{iso}^{calc}, for different radical densities of PTMA, are compared against experimentally derived *g* values obtained from *in operando* EPR measurements of a PTMA-based organic radical battery. The ML-predicted g_{iso} values, g_{iso}^{pred}, were compared with g_{iso}^{calc} to evaluate the performance of the model. Mean deviations of g_{iso}^{pred} from g_{iso}^{calc} were found to be on the order of 0.0001. Furthermore, a performance evaluation on test structures from a separate MD trajectory indicated that the model is sensitive to the radical density and efficiently learns to predict g_{iso} values even for radical densities that were not part of the training data set. Since our trained model can reproduce the changes in g_{iso} along the MD trajectory and is sensitive to the extent of equilibration of the polymer structure, it is a promising alternative to computationally more expensive DFT methods, particularly for large systems that cannot be easily represented by a smaller model system.



INTRODUCTION

Electron paramagnetic resonance (EPR) spectroscopy, a technique to characterize systems with unpaired electrons, is highly effective in structure elucidation and mechanistic investigations when coupled with theoretical modeling.[1,2] While EPR selectively probes unpaired electrons and can be used to gain insights into the electronic structure of spin systems, electronic structure methods such as density functional theory (DFT) enable the calculation of EPR spectroscopic observables and confirm experimental inferences made by using EPR methods. In the case of small molecules, DFT techniques are a popular choice for accurate calculations of physical and chemical properties at moderate computational costs.[3,4]

A combination of DFT and molecular dynamics (MD) is frequently employed for simulating EPR spectra of dynamic systems, and EPR parameters calculated using DFT methods often serve as initial values for least-squares fitting of EPR spectra.[5] In order to fully capture the evolution of a specific property along an MD trajectory, a large number of DFT calculations are required. For larger systems, DFT calculations can become considerably more expensive, requiring high-performance computing systems to afford such calculations. As a viable alternative, machine learning (ML) methods have seen a recent surge in application, especially in materials[6] and chemical sciences.[7,8] ML algorithms can be used to efficiently learn structure−property correlations and make predictions for unknown structures with comparable accuracy as DFT at a fraction of the computational time. Various physical and chemical properties including chemical shifts,[9] ground-state energies,[10] or redox potentials[11] have been targeted using ML methods.

A key parameter derived from EPR spectra is the *g* value, which describes the interaction of the electron spin with the applied magnetic field. The observed *g* value deviates from the free electron *g* value (g_e = 2.00232) mainly due to spin−orbit coupling, which is a relativistic effect arising from the interaction of spin and orbital angular momentum. Spin−orbit coupling is more pronounced for heavier elements. Transition metal ions can show large deviations from g_e, while organic radicals show smaller deviations in comparison.[12] Therefore, the *g*-shift is characteristic of the chemical identity of the system. In the solid state and frozen solutions, the value

Received: November 13, 2023
Revised: February 26, 2024
Accepted: February 26, 2024
Published: March 8, 2024

of g is dependent on the orientation of the molecule with respect to the externally applied magnetic field, leading to g anisotropy described by a $\mathbf{g}$ tensor with three principal values: g_1, g_2, and g_3.[13,14] In the liquid state, the anisotropy is averaged out due to fast tumbling of molecules, and an isotropic g value, $g_{iso} = (g_1 + g_2 + g_3)/3$, is obtained. As $\mathbf{g}$ is related to the underlying molecular geometry, g values can be used to investigate electronic distributions in molecules. This is particularly relevant in the case of transition metal complexes, where an analysis of the $\mathbf{g}$-tensor components can be used to study geometrical distortions.[15] For organic radicals, g values can be a sensitive probe for the identification of the radical center and elucidation of structural changes in its environment.[16−18] The magnitude of g value shifts as a result of environmental changes in organic radicals can be minimal, usually of the order of 10^{-4}, and high-field EPR measurements are used to gain the required spectral resolution.[17,19]

In systems with high spin concentration and small interspin distances, spin−spin interactions such as the Heisenberg spin exchange or dipole−dipole interactions considerably affect the EPR spectrum. As a consequence of strong exchange, the coalescence of spectral features may occur and the EPR spectrum is mainly characterized by g_{iso} and the line width.[20] Prominent examples of such systems include organic radical polymers (ORPs), where a high radical concentration and closely spaced radicals are prerequisites for energy storage applications, leading to an exchange narrowed EPR line. ORPs are instrumental in the pursuit of more sustainable energy storage technologies.[21] They consist of pendant radical moieties as repeat units. These radical centers have unpaired electrons, which impart electrochemical activity to the polymer by undergoing redox reactions. The polymer structure used in this work, poly(2,2,6,6-tetramethylpiperidinyloxy-4-yl methacrylate) (PTMA), consists of 2,2,6,6-tetramethylpiperidinyloxyl (TEMPO) radicals as the redox unit with a methacrylate backbone. Owing to their electrochemical properties, ORPs find extensive use in emerging battery technologies such as organic radical batteries (ORBs).[22] The redox properties can be tuned through chemical synthesis, enabling the construction of all-organic batteries,[23−25] devoid of toxic metals. As the radicals cause paramagnetic properties, ORPs constitute a class of compounds that can be investigated using EPR.

EPR $\mathbf{g}$ tensors can be calculated at the DFT level.[26−29] The $\mathbf{g}$ matrix is obtained through the calculation of three main contributions to the g shift consisting of a relativistic mass correction term (Δg^{RMC}),[30] a diamagnetic gauge correction term (Δg^{DGC}),[31] and a term related to orbital Zeeman and spin−orbit coupling interactions ($\Delta g^{OZ/SOC}$).[26,32] The first two terms, Δg^{RMC} and Δg^{DGC}, can be obtained from the spin density. At the DFT level, $\Delta g^{OZ/SOC}$ can be obtained through a solution of coupled-perturbed self-consistent field (CP-SCF) equations, which is also the approach used in the *ORCA* software package.[26,33]

We recently reported the use of the g value as a parameter for method validation of MD simulations of radical polymers.[34] The evolution of g was monitored by calculating g using DFT for different time frames of the MD trajectory. Since g can be experimentally verified, characteristics of the simulated polymer can be tuned to match realistic sample conditions. For instance, experimental g can be compared with calculated g to judge the simulation time scales needed to obtain an equilibrated polymer structure. Furthermore, due to the sensitivity of g toward the molecular structure, features like the minimum chain length of the simulated polymer can be optimized.

DFT-based geometry optimization and $\mathbf{g}$-tensor calculations on large disordered systems such as polymers are computationally demanding. While electronic structure methods that scale linearly with the number N of atoms in a system have been reported,[35,36] computational costs of most DFT calculations scale with N^3 for larger system sizes.[37,38] In the case of $\mathbf{g}$-tensor calculations, SCF procedures scale with N^2.[26] Adequate theoretical representations of radical polymers often require several monomers, which increase the system size ($N >$ 200 for one polymer molecule), making the application of DFT methods less feasible. Additionally, experimental investigations of the cycling stability of ORBs indicate the importance of using cross-linked polymers, and, therefore, simulation systems need to include and account for additional cross-linking moieties, increasing the complexity of the simulated system further.[34]

Using *in operando* EPR techniques, evolution of the active material as a function of the state of charge can be studied.[39−41] Here, the g value serves as a parameter that can be theoretically computed and experimentally verified to substantiate the required complexity of the simulated system and its similarity to the experimentally investigated states of charge. While simulation of battery systems with realistic complexity for various states of charge is attainable using MD, DFT implementations result in inferior scalability. Therefore, a computationally cheaper and scalable approach to predict g values for larger and more complex simulation systems would be desirable.

ML approaches provide an alternative to bypass the computational expense of electronic structure methods such as DFT. For instance, ML methods have been applied to learn density functionals itself, with the aim of avoiding the calculation of Kohn−Sham equations.[42] Machine learnability of properties calculated using more accurate and computationally demanding electronic structure methods such as coupled cluster has also been reported.[43] Learning algorithms that utilize neural networks usually require large data sets for training, and the learning process can be computationally expensive. In the case of size-limited data sets, ML algorithms such as regression trees, Gaussian process regression (GPR), support vector regression (SVR), or kernel ridge regression are more suitable.[44−46] For instance, applicability of tree-based algorithms for pK_a predictions in proteins[47] and GPR for atomistic properties[48,49] has been successfully demonstrated.

A key advantage of complementing electronic structure methods with ML techniques is speeding up the computation of properties and not limiting their calculation to small-sized systems. Within the field of magnetic resonance, ML is an emerging tool with similar aims. In particular, applying ML for nuclear magnetic resonance (NMR) chemical shift predictions from molecular structures has gained widespread use.[9,50] ML methods also find utility in EPR data analysis, mainly in the subfield of hyperfine spectroscopy[51] and dipolar spectroscopy.[52] More recently, neighborhood component analysis, an ML algorithm, was utilized to quantify the importance of structural parameters in determining electron−nuclear hyperfine interaction tensors.[53] Although analogous to chemical shifts, to the best of our knowledge, ML approaches to predict EPR g values have not been attempted yet.

In this work, an ML model is developed to predict g values from molecular structures of PTMA. The model is trained on

https://doi.org/10.1021/acs.jctc.3c01252
J. Chem. Theory Comput. 2024, 20, 2592−2604

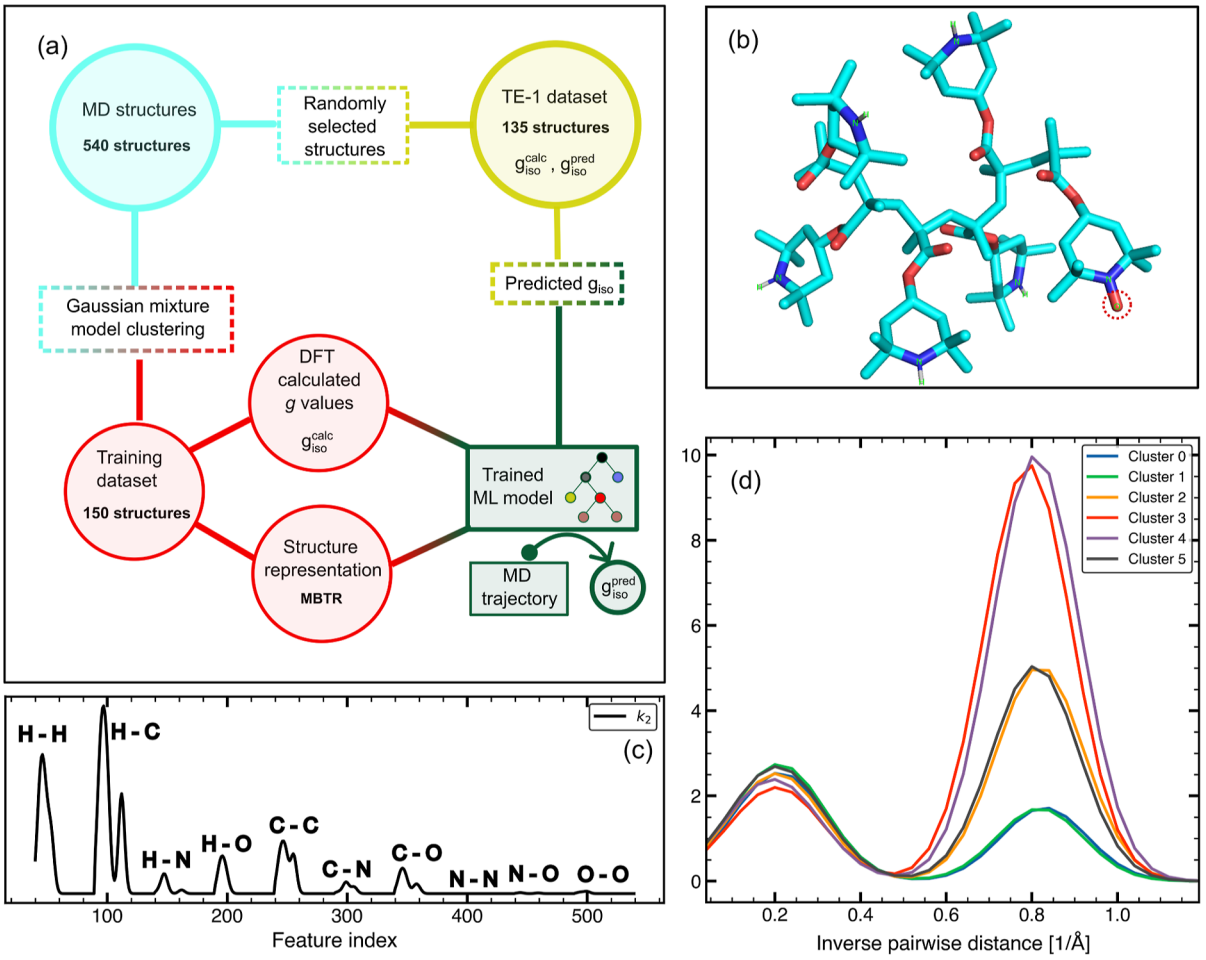


Figure 1. (a) Schematic representation of the workflow used for machine learning g_{iso}. (b) Ball and stick model of PTMA-1, with six monomer units out of which one is a radical center (circled with a red dotted line). Oxygen atoms are colored red, nitrogen atoms are colored blue, and carbon atoms are colored cyan. (c) Schematic representation of the k_2 term of MBTR for a PTMA-1 molecule. (d) Section of the N−O region of MBTR showing structural differences in cluster centers obtained using GMM. The distribution centered at an inverse pairwise distance of ≈0.8 $Å^{-1}$ corresponds to the N−O bond length, and the distribution centered at ≈0.2 $Å^{-1}$ corresponds to long-range N−O interactions. The distributions are weighted exponentially to give more weight to short-range interactions.

the value to be predicted, g_{iso}, and the corresponding structural characteristics of the polymer as the features. The dynamic evolution of g along the MD trajectory, which is correlated to the underlying molecular structure, is exploited to train the model with the aim of reproducing the evolution of g for unknown radical densities or, alternatively, for different states of charge of an ORB. The performance of the model is evaluated on test data sets, and the dependence of model performance on different molecular descriptors is discussed. Finally, the trained model is applied to an unknown MD trajectory to test its ability to completely predict the evolution of g along the trajectory. The model is also studied with respect to the quality of interpolation to unknown radical densities. The importance of specific features that determine the predictions of the model is evaluated and discussed in correlation with the molecular structure of the polymer.

WORKFLOW

The workflow followed in this work is depicted in Figure 1a. First, PTMA polymer molecules are dynamically evolved in the presence of an electrolyte using classical MD (see "MD Simulations" in the Methods section) to simulate an organic cathode. For MD simulations, the polymer is represented by using a linear polymer structure made of six monomer units. Structures with different radical densities were generated by varying the number of monomers that are radicals. The notation used in this work is PTMA-X, where X denotes the number of monomers which are radicals. For instance, PTMA-6 represents 100% radical density, i.e., all six monomers are 2,2,6,6-tetramethylpiperidinyloxy-4-yl methacrylate (TEMPO methacrylate) radicals, while PTMA-1 corresponds to a structure where only one out of six monomers is a radical and the other five monomer units are diamagnetic 2,2,6,6-tetramethylpiperidin-4-yl methacrylate groups (see Figure 1b). From separate MD trajectories of PTMA-1, PTMA-3, and PTMA-6, polymer structures from different time frames (see "Data Set Generation" in the Methods section) are extracted to generate the whole structural data set (**WSD**).

For ML, input data need to be transformed into suitable features, which can be used to train the ML model. To predict structure-dependent properties, structural features of the molecule, such as interatomic distances and bond angles, can be utilized as features.[54] To encode such features and represent the structure, the input data are transformed using a molecular descriptor. Molecular descriptors encode either a local atomic environment or the whole structure, depending on the

https://doi.org/10.1021/acs.jctc.3c01252
J. Chem. Theory Comput. 2024, 20, 2592−2604

property to be predicted. Local descriptors, such as smooth overlap of atomic positions (SOAP),[55,56] encode local atomic environments, while global descriptors, such as many-body tensor representation (MBTR),[56,57] encode the whole structure. The representations can be further tuned by user-defined and system-dependent parameters to capture only relevant atomistic interactions. In addition to the aforementioned descriptors, a third set of features, generated from bond lengths, bond angles, and dihedral angles [distances−angles−dihedrals (DAD)] extracted from the polymer structure, is also utilized for representing the molecular structure in this work. The parameterization and construction of molecular descriptors are described under "Molecular Representation" in the Methods section.

To ensure structural diversity in the training data set (**TR**), an optimum sampling of the configurational space of the polymer molecule is required. Clustering algorithms such as the Gaussian mixture model (GMM) clustering can be used to find clusters in the data set and sample configurationally diverse PTMA structures.[58−61] GMM assumes that the data to be clustered originate from a mixture of multivariate Gaussian distributions with unknown means and covariances. GMM fits a specified number of distributions, which correspond to clusters in the data set and optimize the values of mean and covariance through the expectation−maximization algorithm.[62] The cluster centers, or the mean of Gaussian distributions, are initialized randomly from structures in the data set itself and iteratively updated to maximize the likelihood of the data with respect to the distributions. To generate the **TR**, the atomic coordinates of structures in **WSD** are transformed using a suitable molecular descriptor. By clustering the structures in **WSD** multiple times with random initial clusters and selecting the structure at the center of each cluster on each run, the **TR** is obtained. To illustrate that this approach finds structurally diverse structures, the N−O distance regions of the MBTR output corresponding to six cluster centers from a single run of the clustering algorithm are shown in Figure 1d.

DFT calculations for obtaining g_{iso} need to be done only for the structures selected by the clustering algorithm, thereby reducing the computational cost of the ML workflow. From the remaining structures of **WSD**, which are not included in the **TR** data set, a subset of structures are randomly selected to generate a test data set **TE-1** for performance evaluation of the trained model. The structures in **TR** represented using a molecular descriptor and the corresponding DFT-derived g_{iso} (g_{iso}^{calc}) are used for training the model.

To select an appropriate learning method, various models trained using different types of learning algorithms were evaluated with **TR** using cross-validated scoring [see Section A1 and Figure S1 in the Supporting Information for a description of cross-validation]. The tested methods are chosen based on their known applicability to predict structure-dependent properties or for generating baseline models that warrant the need for more complex models. For instance, SVR, GPR, and tree-based ensemble methods have previously been demonstrated for prediction of properties such as NMR chemical shifts, which are analogous to g values.[9,63,64] The performance of each trained model was then evaluated through cross-validated scoring with the **TR** data set represented using SOAP, MBTR, and DAD (see Section A2 in the Supporting Information). The model molecular descriptor pair with the best and most consistent predictive accuracy, based on the standard deviation of the error metric, is selected as the final model. The final model is then applied to structures derived from an unknown MD trajectory for further evaluation.

METHODS

MD Simulations. For each of the three radical densities studied, PTMA-1, PTMA-3, and PTMA-6, the following classical MD simulation[65] was used. Hereafter, this simulation is referred to as MD-1. Six monomers were used to represent each of the PTMA polymers and a total of 24 such polymers were used to mimic an organic electrode.[34] As an electrolyte solution, 1064 ethylene carbonate (EC), 2100 ethyl methyl carbonate (EMC), and 300 $LiPF_6$ molecules were chosen in a $10 \times 10 \times 10$ nm^3 simulation cell, with periodic boundaries along all three Cartesian coordinates. Partial charges for EC, EMC, and PF_6^- were taken from the previous work.[66−69] The initial configurations ($t = 0$) were constructed using PACKMOL,[70] which avoids repulsive potentials by keeping a safe interatomic distance. A two-step NPT ensemble process was undertaken for all MD simulations using GROMACS 2019.[71] Initially, a 2 ns initialization step with a 0.5 fs time step ensured system stabilization, where the Berendsen thermostat and barostat maintained the temperature at 298.15 K and pressure at 100 bar with 1.0 ps time constants.[65] Subsequent to the initialization step, an equilibration step of 20 ns with a reduced 1 fs time step employed the Nosé−Hoover thermostat and Parrinello−Rahman barostat to regulate the temperature and reduce the reference pressure to 1 bar, keeping time constants constant at 1.0 ps. Throughout both steps, Coulombic and Lennard−Jones interactions were handled via a particle−particle-mesh solver with a consistent cutoff of 1.2 nm. The OPLS all-atom force field[72] was used for all MD simulations. The atomic site charges on paramagnetic and diamagnetic repeating units of PTMA were calculated from the electrostatic potential (ESP) fit (see Figure S5 in the Supporting Information). Gaussian16[73] was used to calculate the ESP charges using the MP2 theory[74] with a pVDZ basis set.[75] The structures from the MD-1 trajectory were used for training the model (see the "Data Set Generation" section).

For additional model evaluation, separate MD simulations (MD-2), with only one polymer chain in the simulation box, were used for generating PTMA-1, PTMA-2, PTMA-3, PTMA-4, and PTMA-6 structures. Other MD parameters were kept the same as for MD-1. The structures from MD-2 were solely used for generating test data sets (see the "Data Set Generation" section).

DFT Calculations. DFT computations of g values were conducted using ORCAv. 5.0.2[76] and PTMA polymer structures obtained from different time frames of the MD trajectory. The **g**-tensor origin was set to the center of spin density and calculations were done using the unrestricted Kohn−Sham formalism, with the B3LYP functional and EPR-II[77] basis set. Automatic generation of auxiliary basis sets was used for all calculations.[78] Typical calculation times ranged from 1 to 3 h per structure while running in parallel on 12 cores. Optimization of the DFT calculation protocol for TEMPO methacrylate and PTMA is described elsewhere.[34]

Data Set Generation. To generate the **WSD**, a total of 540 structures were sourced from separate MD trajectories of PTMA-1, PTMA-3, and PTMA-6. From the initialization step of MD-1, seven logarithmically spaced time frames were sampled from 0.0005 to 500 ps. From the equilibration step of MD-1 ranging from 2000 to 20,000 ps, structures were

https://doi.org/10.1021/acs.jctc.3c01252
J. Chem. Theory Comput. 2024, 20, 2592−2604

sampled linearly in steps of 2000 ps. As structures from the same time frame show variation in the DFT-derived g_{iso} (g_{iso}^{calc}), a total of 10 structures per time frame were extracted. In addition, 10 structures for each radical density from the start of the MD trajectory (t = 0) were also added to **WSD**. The number of PTMA-1, PTMA-3, and PTMA-6 structures is equal in **WSD**. To generate the **TR**, GMM as implemented in scikit-learn v. 1.2.2[79] was applied to structures in **WSD**, with each structure represented using a molecular descriptor (see the "Molecular Representation" section). For GMM, initialization of clusters was done randomly from the data set. Upon convergence of the clustering algorithm, the structure at the center of each cluster (or the mean of each Gaussian distribution) was added to the **TR**. Due to the randomness in selecting the initial clusters, 50 runs of the algorithm were done to obtain 150 unique structures that form the **TR**. The test data set **TE-1** was generated by randomly selecting 135 structures from the remaining structures of **WSD**. Another test data set **TE-2** was generated by sourcing structures from the MD trajectory of MD-2, consisting of equal numbers of PTMA-1, PTMA-2, PTMA-3, PTMA-4, and PTMA-6 structures.

Molecular Representation. Chemical structures in the XYZ file format were read using the Python library atomic simulation environment (ASE) v. 3.22.1.[80] For implementation of SOAP[55] and MBTR,[54] the DScribe library[56] (v. 1.2.2) was used. In the SOAP formalism, atomic positions were represented by using Gaussian functions. This allows for structural representation in terms of atomic neighbor density around a central atom given by

$$\rho(\boldsymbol{r}) = \sum_i e^{-\frac{1}{2\sigma^2}|\boldsymbol{r}-\boldsymbol{R}_i|^2} \tag{1}$$

where $\boldsymbol{r}$ is the position in the space of the central atom, $\boldsymbol{R}_i$ is the position of a neighboring atom, and summation i runs over all atoms within a specified cutoff distance from the central atom. The atomic density was calculated for each atom in the molecular structure and was expanded using a combination of spherical harmonics and radial basis functions (RBFs).[55] The final SOAP output was constructed as a power spectrum of the expanded atomic density.[56]

In the MBTR, a geometry function was used to transform a configuration of k atoms into a single value. Commonly used geometric functions for k = 1, k = 2, and k = 3 configurations corresponded to atomic numbers (Z), inverse pairwise distances $|\boldsymbol{R}_1 - \boldsymbol{R}_2|^{-1}$, and angles between three atoms $\angle(\boldsymbol{R}_1 - \boldsymbol{R}_2, \boldsymbol{R}_3 - \boldsymbol{R}_2)$, respectively. The values obtained from the geometric functions were then converted into distributions using the Gaussian kernel density estimation. The final MBTR output consisted of concatenated Gaussian distributions corresponding to all possible combinations of k atoms and was weighted according to the number of occurrences of a particular combination in the molecular structure.[56] As an example, Figure 1c shows the k_2 term of the MBTR output for a molecule of PTMA-1 (see Figure 1b), depicting the inverse distances between atom pairs of the PTMA oligomer.

In the case of SOAP and MBTR, cutoff distances (see Table S4) and number of features (see Table S5) were optimized using the **TR** data set and cross-validated scoring (see Section A6 in the Supporting Information). Parameters for feature vectors used to train the final model are described below. In the case of SOAP, the cutoff distances were set to 10 Å, and spherical Gaussian-type orbitals were used as RBFs. The number of RBFs and maximum degree of spherical harmonics was set to 8. For MBTR, geometric functions for k_1, k_2, and k_3 terms corresponded to atomic numbers, inverse pairwise distances in units of Å^{-1}, and angles between three atoms in degrees, respectively. The bounds for the k_1 term were set to 0 and 10 with the number of discretization points set to 10. For k_2, a distance range of 0.5−25 Å (inverse distance range of 0.04−2 Å^{-1}) was used and the number of discretization points was set to 50. For the k_3 term, the number of discretization points was set to 180. In the case of k_2 and k_3, the distributions were weighted exponentially as a function of distance to give more importance to structural properties corresponding to closely spaced atoms. A distance cutoff of 10 and 5 Å was used for weighting the k_2 and k_3 terms, respectively. The Gaussian smoothing width (σ) was set to 0.01 for the k_1 term, 0.1 for the k_2 term, and 4 for the k_3 term. The DAD molecular representation was built using bond lengths (Å) between two atoms, bond angles (°) made by three atoms, and dihedral angles (°) between two atomic planes. Bond lengths, angles, and dihedral angles were extracted using ASE. For all structural features, only bonded atoms within a cutoff distance of 1.5 Å were considered. The values were sorted according to the respective atomic combinations and concatenated to form the final DAD feature vector.

Model Building. The ML algorithm for the final model was selected using cross-validated scoring with the **TR** data set (see Section A2 and Figure S2 in the Supporting Information). The main types of considered methods were SVR,[81] GPR,[9,48] regression trees,[82] ensemble methods, and linear regression. In the case of ensemble methods, averaging[44,83] and boosting[84,85] techniques were tested.

The final model based on regression trees (see Section A3 in the Supporting Information for a theoretical description), which is used in this work, was built as follows. A regression model based on the extremely randomized trees (ERT) method[46] was generated using scikit-learn v. 1.2.2[79] (ExtraTreesRegressor in *scikit-learn*). The model was trained on the **TR** data set. Parameters that affect the learning process, known as hyperparameters, were optimized through a bound constrained and exhaustive grid search (GridSearchCV in *scikit-learn*). The root mean squared error (RMSE) was minimized during the optimization. The grid search bounds and optimized hyperparameters are listed in Table S1 in the Supporting Information. Hyperparameter optimization was also attempted using Optuna[86] (see Section A4 in the Supporting Information) and the optimized parameters are listed in Table S2. For the final model, parameters obtained from the grid search were used, based on the performance metrics (see Table S3). The ERT model was always initialized with a random state of 1 for reproducibility. The performance of the model on the test data sets TE-1 and TE-2 was quantified using the coefficient of determination (R^2), mean absolute error (MAE), and RMSE. A mathematical description of the error metrics is given in Section A5 of the Supporting Information. The training time of the final model using the **TR** data set (represented using MBTR) with the optimized hyperparameters was 4.97 ± 16.6 ms while running on a single core (Apple M1 processor) and 926 ± 13.7 ms when parallelized on 8 cores (Apple M1 processor). The time required for prediction was 3.58 ms ± 67.2 μs (single core). The reported times are mean ± standard deviation of 7 runs.

https://doi.org/10.1021/acs.jctc.3c01252
J. Chem. Theory Comput. 2024, 20, 2592−2604

RESULTS AND DISCUSSION

Evolution of Calculated g_{iso}. Figure 2 shows the DFT-calculated g values (g_{iso}^{calc}) for polymer structures extracted from

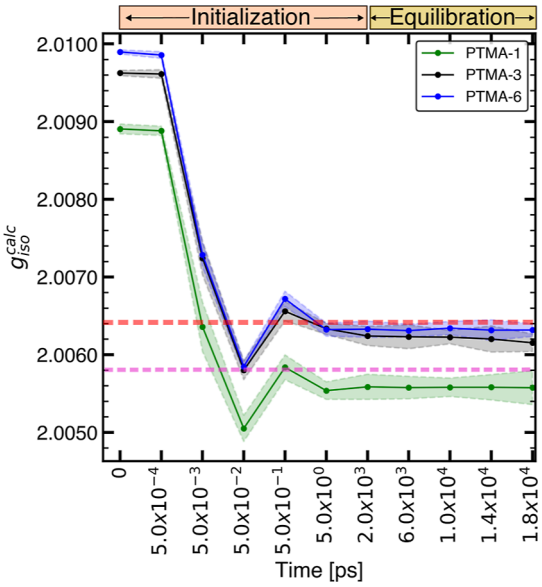


Figure 2. Evolution of g_{iso}^{calc} along the MD-1 trajectory. Each data point is the average of g_{iso}^{calc} for 10 individual structures. Shaded regions denote the standard deviation. Before the equilibration run starting at 2000 ps, an initialization step was done. Experimental g_{iso} for PTMA (red) and its monomer (magenta) in the solution state equal to 2.0064 and 2.0058, respectively, and are denoted by dashed lines.

different time frames of the MD trajectory. Each data point represents the mean ($\overline{g}_{iso}^{calc}$) of g_{iso}^{calc} of 10 randomly selected structures for a specific time frame. The structures differ in their radical density, i.e., the number of monomer units that are radicals. The experimental analogue for the PTMA-6 (100% radical density) is a PTMA polymer in a state of high radical density, such that radical−radical interactions are significant. In the case of PTMA-1, $\overline{g}_{iso}^{calc}$ is compared to the experimental g value obtained from a dilute solution of TEMPO methacrylate (PTMA monomer), with negligible radical−radical interactions. As an additional experimental reference, experimental g values of PTMA for transient states of charge are obtained from the *in operando* EPR measurement of an ORB with PTMA as the active material (see Section A10 in the Supporting Information). The *in operando* EPR experiments show that changes in g are reversible during battery cycling and are correlated to the radical density (see Figure S6a,b in the Supporting Information).

Experimental g values (g_{iso}^{exp}) of the PTMA polymer and TEMPO-methacrylate solutions were reported to be 2.0064 (dashed red line in Figure 2) and 2.0058 (dashed magenta line in Figure 2).[34] For a pristine ORB, g_{iso}^{exp} of PTMA was found to be 2.0065 (see Figure S6a in the Supporting Information). For PTMA-6, g_{iso}^{calc} is in good agreement with both corresponding experimental references, with $|g_{iso}^{exp} - \overline{g}_{iso}^{calc}|$ being 0.0001 in the case of the polymer solution and 0.0002 in the case of PTMA with high radical density in the ORB. At an intermediate radical density of PTMA in the ORB, g_{iso}^{exp} already approaches values which are predicted by DFT for PTMA-1, a simulated structure with low radical density. The comparison of g_{iso}^{calc} corresponding to PTMA-1 with g_{iso}^{exp} of a dilute solution of TEMPO methacrylate shows a difference of 0.0003.

For both experimental references, a better agreement between g_{iso}^{exp} and g_{iso}^{calc} is observed at high radical density in comparison to that at low radical density. The larger deviation between g_{iso}^{exp} at lower radical density and g_{iso}^{calc} of PTMA-1 may be explained by considering the structural and conformational differences between the experimental and simulated structures. A dilute solution of nitroxide radicals does not include the structural effects of the extended oligomer chain. In the simulated structure of PTMA-1, the radical species is restricted to the oligomer backbone, while in a dilute solution of TEMPO methacrylate, the radical undergoes isotropic motion. In the case of PTMA in an ORB, the disagreement between g_{iso}^{calc} of PTMA-1 and g_{iso}^{exp} for a state of low radical density of the ORB may stem from interactions between PTMA and other battery constituents, which are more effective when intramolecular radical−radical interactions are negligible. It is known that conductive additives used in organic cathodes affect experimental EPR parameters and battery performance.[87] However, the current complexity of the simulated system does not account for such interactions. Another reason could be long-range radical−radical interactions which are not captured in the DFT calculations with individual polymer chains but are expected to be present even at low and intermediate radical densities as the polymer prefers to maximize radical−radical interactions.[34,88,89] On the other hand, the better agreement at higher radical densities also indicates that at high radical density, for any particular redox unit, radical−radical interactions are adequately represented by the simulated structure consisting of only radicals. Inclusion of multiple, longer polymer chains and modeling of conductive additives in the simulated system may improve the agreement also for low and intermediate radical densities. However, EPR parameters of such complex systems at transient states of charge might not be accessible by DFT methods due to the high computational cost.

The PTMA structure at time $t = 0$ for each radical density corresponds to the structure before equilibrium and shows a large deviation from g_{iso}^{exp}. However, the extent of deviation seems to be dependent on the radical density, providing a measurable g value difference between the structures before equilibrium for each radical density. Furthermore, the structure at $t = 0$ serves as an important evaluation criterion for the ML model to assess its ability to differentiate between radical densities. Timeframes up to 2000 ps correspond to the initialization step. A rapid convergence of $\overline{g}_{iso}^{calc}$ to g_{iso}^{exp} is observed as the simulated system stabilizes. Changes in $\overline{g}_{iso}^{calc}$ are minimal beyond 50 ps, indicating an equilibrated system.

DFT-derived g_{iso}^{calc} values allow for a differentiation of the polymer structures through two aspects. First, structures before equilibrium and structures after equilibrium show significant differences in g_{iso}^{calc}, and second, g_{iso}^{calc} changes with respect to the radical density. For an ML model to be a viable alternative, the model should also learn to differentiate polymer structures based on the two aforementioned aspects. Additionally, the evolution of g_{iso}^{calc} along the MD trajectory can only be reproduced if the model learns to correlate the magnitude of change in g_{iso}^{calc} to the extent of structural changes. Despite structural differences, PTMA-3 and PTMA-6 showed small differences in converged values for $\overline{g}_{iso}^{calc}$, while a larger contrast

https://doi.org/10.1021/acs.jctc.3c01252
J. Chem. Theory Comput. 2024, 20, 2592−2604

is obtained when PTMA-1 is compared to PTMA-3 and PTMA-6.

By including PTMA-3 in the **TR**, the trained model can be expected to learn smaller g_{iso} changes or equivalently subtle structural deviations, which occur as a result of a change in the radical density. This should allow for more precise g_{iso} interpolations for radical densities that lie within the limits of **TR** but are not included in the **TR**.

Performance Evaluation and the Analysis of Regression Tree Models. Among the tested ML methods, ensemble methods which use a combination of multiple models showed the best performance with mean deviations of g_{iso}^{calc} from predicted g_{iso} (g_{iso}^{pred}) ranging from 0.0001 to 0.0004 (see Figure S2 in the Supporting Information). In comparison, kernel-based methods, GPR and SVR, showed larger mean deviations ranging from 0.0002 to 0.0006. Overall, models trained on MBTR feature vectors showed better predictive accuracy in comparison to SOAP and DAD feature vectors. Through cross-validation, the spread of MAE and RMSE scores also suggests that the models trained on MBTR perform consistently better on different folds of the **TR** data set. With MBTR feature vectors, even simpler models such as linear regression show RMSE values of the order of 0.0004, while the model performs considerably worse in the case of DAD. This is likely due to the nonlinear mapping of structural properties in the case of MBTR, which was also observed previously for ML-based predictions of exchange spin coupling.[49] In DAD, the values of angles and distances are used without further transformation, and a linear relationship between the structural properties and g_{iso} may not be present. However, in the case of kernel-based methods and ensemble methods, which can handle nonlinear data, models trained on DAD show improved performance. A combination of ERT and MBTR showed the best performance with the lowest MAE and RMSE values (see Section A2 in the Supporting Information).

Table 1 compares the cross-validated error metrics for the ERT model in the case of three different molecular descriptors investigated in this work. ERT models trained on global descriptors perform considerably better than in the case of the ERT model trained on SOAP feature vectors. This indicates that in the case of the polymer structure, the property to be learned, g_{iso}, is affected by the structural features of the whole structure rather than a localized region. This observation is in agreement with CW-EPR experiments using PTMA, where polymer and monomer samples show variations in the observed g value. Furthermore, as shown in the previous section, DFT calculations using polymer structures with different radical densities also show structure-dependent changes in g_{iso}^{calc}. *In operando* EPR results (see Section A10 in the Supporting Information) further confirm that the g_{iso} value of the active material, PTMA, is influenced by the change in the radical density (see Figure S6). As changes in the radical density can be accompanied by changes in polymer conformation,[34,88,89] descriptors which encode the global polymer structure, such as MBTR and DAD, fare better in g value predictions. The performance of the ERT model was also studied with respect to the cutoff distance used for SOAP and MBTR. For SOAP, a cutoff distance of 10 Å was found to be optimal. In the case of MBTR, slight improvements in RMSE were observed with larger cutoff distances (see Section A6 in the Supporting Information). A cutoff distance of 10 Å was used for both SOAP and MBTR based on the RMSE (see Table S4). As the distance between two adjacent monomer units on the oligomer chain lies in the range of 7−10 Å, a cutoff distance of 10 Å should capture most of the intramolecular interactions between the monomer units.

Table 1. Cross-Validated Performance Metrics Obtained for Different Molecular Descriptors Using the TR Data Set

molecular descriptor	R^2	MAE [$\times 10^{-4}$]	RMSE [$\times 10^{-4}$]
SOAP	0.87 (0.085)	1.91 (0.37)	2.59 (0.68)
MBTR	0.95 (0.019)	1.25 (0.14)	1.58 (0.20)
DAD	0.92 (0.034)	1.53 (0.25)	2.07 (0.45)

Among the two global descriptors, the performance was better for the model trained on MBTR feature vectors. While DAD also encodes the whole structure in the form of bond lengths, bond angles, and dihedral angles, MBTR likely captures long-range interactions of adjacent radical units better due to the larger cutoff distance. This aspect becomes especially important when interpolating predictions for different radical densities, where the presence or absence of a neighboring radical moiety affects g_{iso}^{calc}. However, prompted by similar performance metrics, ERT models trained on both MBTR and DAD were evaluated using the TE-1 test data set.

Figure 3a−d summarizes the dependence of model performance on MBTR and DAD. Hereafter, the ERT models trained on MBTR and DAD features are referred to as ERT-MBTR and ERT-DAD, respectively. The error metrics given in Table 2 indicate 27% (based on MAE) better performance for ERT-MBTR. In comparison to ERT-DAD, a narrower spread (see Figure 3b) of $g_{iso}^{calc} - g_{iso}^{pred}$ and a lower RMSE for ERT-MBTR suggest that the predictions do not deviate heavily from g_{iso}^{calc}. ERT-MBTR shows superior prediction accuracy especially in the g value range of 2.0090−2.0100 (see Figure 3a) when compared to ERT-DAD (see Figure 3c). The structures in this range, which show a significant deviation from the experimental g values, are obtained from the initialization step before the start of the equilibration step (t = 2000 ps). The ability of the model to differentiate between structures before equilibrium and equilibrated structures is essential for application to new MD trajectories. For equilibrated structures in the case of PTMA-6 and PTMA-3, g_{iso}^{calc} lies in the range of 2.0060−2.0066. As differences between structures are minimal in this regime, predictions in agreement with g_{iso}^{calc} require the model to be sensitive to minor changes in the structural features. ERT-MBTR seems be more sensitive to such changes, evident from the better performance compared to ERT-DAD in the corresponding g value range. Model performance of ERT-MBTR remained similar toward all radical densities in **TE-1**. Similar R^2 scores of 0.990, 0.986, and 0.982 were obtained for evaluation using PTMA-1, PTMA-3, and PTMA-6 structures, respectively, indicating that the model was not overfitting features from a specific radical density.

To interpret ERT models further, feature importance scores were analyzed. For simplicity, the importance of each structural feature was summed over all discretization steps for the feature under consideration. For ERT-MBTR, among features which encode pairwise distances, N−O and H−H were found to be important (see Figure 4a).

Dependence on N−O can be rationalized, as PTMA-1, PTMA-3, and PTMA-6 directly differ in the number of N−O bonds. Since MBTR feature vectors employ Gaussian distributions weighted according to the number of occurrences of a particular feature, PTMA-1, PTMA-3, and PTMA-6 with

https://doi.org/10.1021/acs.jctc.3c01252
J. Chem. Theory Comput. 2024, 20, 2592−2604

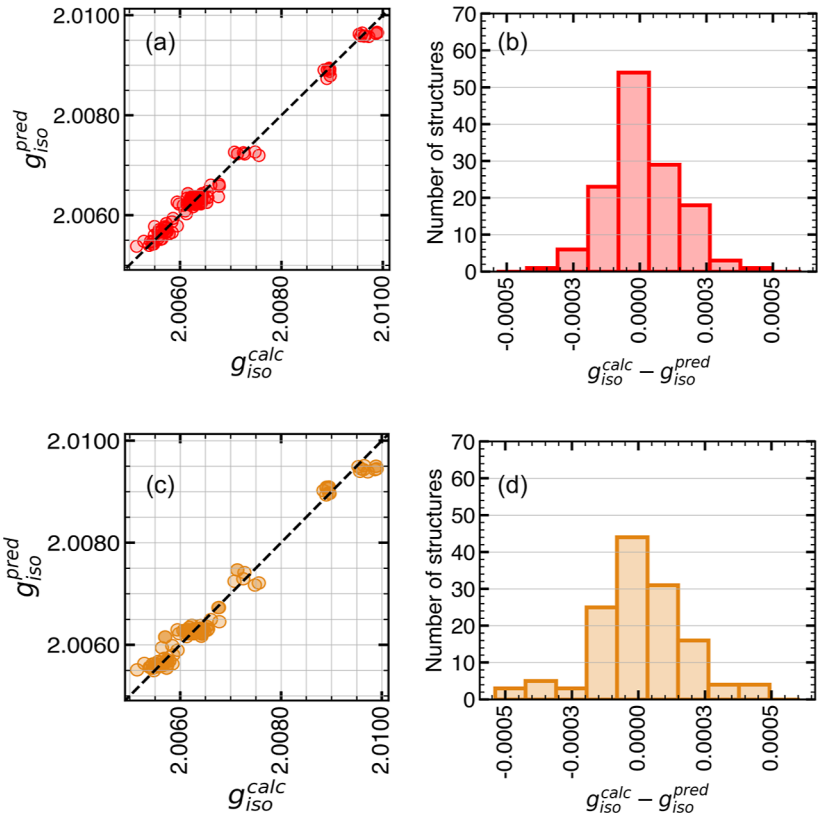


Figure 3. Performance evaluation of the model trained on MBTR and DAD structures using the TE-1 data set. Correlation plot showing g_{iso}^{calc} vs g_{iso}^{pred} for TE-1 structures represented using (a) MBTR and (c) DAD descriptors. Histograms showing $g_{iso}^{calc} - g_{iso}^{pred}$ for TE-1 structures represented using (b) MBTR and (d) DAD descriptors.

Table 2. Error Metrics Obtained for Different Molecular Descriptors Using the TE-1 Data Set

molecular descriptor	R^2	MAE [$\times 10^{-4}$]	RMSE [$\times 10^{-4}$]
MBTR	0.989	0.97	1.27
DAD	0.979	1.28	1.72

1, 3, and 6 N−O bonds can be differentiated. However, a similar importance toward the N−H bond, which serves as an equivalent feature to differentiate between PTMA-1, PTMA-3, and PTMA-6, was not found. Note that further dependence on the N−O feature may arise from through-space interactions within the cutoff distance used for MBTR, for instance, between the nitrogen atom and oxygen atoms of the methacrylate backbone (see Figure 1b). An inspection of the MBTR output for the N−O feature in PTMA-1 (see Figure 1d) shows two distributions centered at ≈1.3 Å (0.8 $Å^{-1}$) and ≈5 Å (0.2 $Å^{-1}$). The shorter distance corresponds to the N−O bond length of TEMPO methacrylate radicals. The longer distance corresponds to the distance between the N atom and the carbonyl O atom of the methacrylate branch. For PTMA-6 and PTMA-3, interactions between N atoms and O atoms on adjacent monomer units may also contribute to the importance of the N−O feature. The N−O distance was also found to be an important distance feature in the case of ERT-DAD, along with N−H (see Section A7 and Figure S3a in the Supporting Information). There are significantly more H−H interaction pairs within the cutoff distance than N−O interactions;

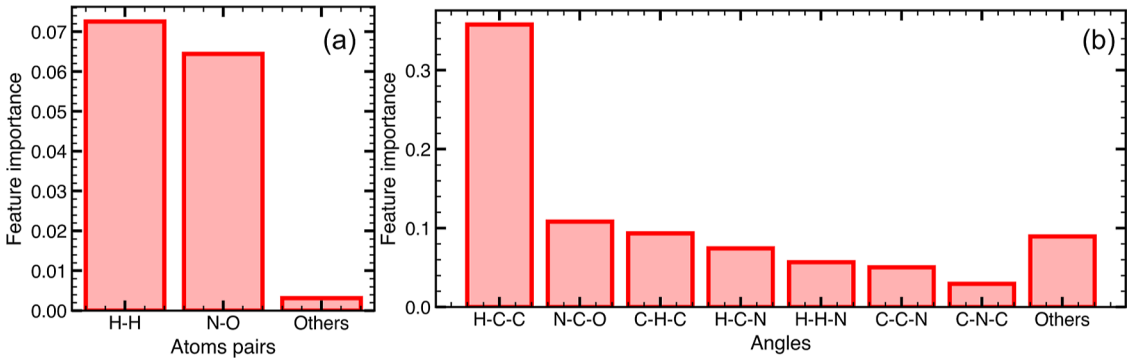


Figure 4. Feature importance scores for ERT-MBTR for (a) distance features and (b) angle features. Angles in "Others" have importance scores less than 0.027.

https://doi.org/10.1021/acs.jctc.3c01252
J. Chem. Theory Comput. 2024, 20, 2592−2604

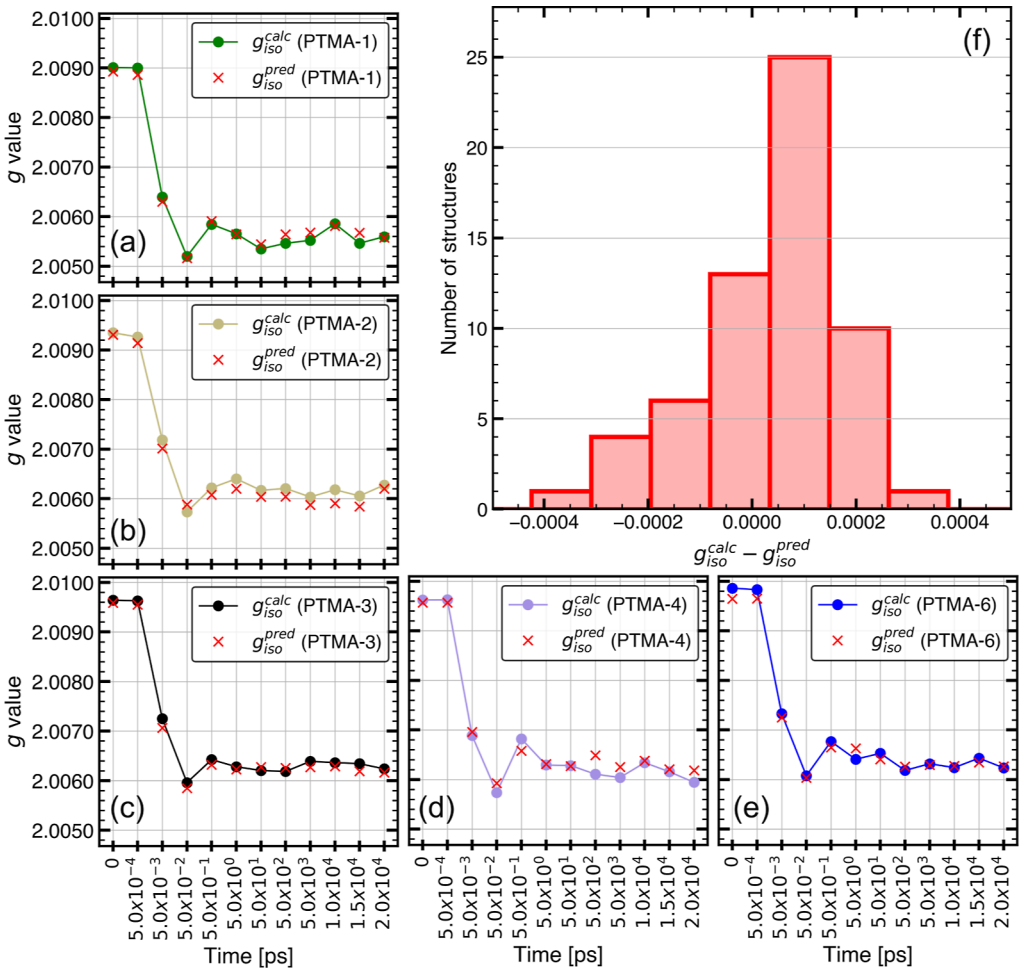


Figure 5. Prediction of g_{iso} from structures in the TE-2 data set. (a) Comparison of g_{iso}^{pred} with g_{iso}^{calc} for TE-2 structures with radical densities of (a) PTMA-1, (b) PTMA-2, (c) PTMA-3, (d) PTMA-4, and (e) PTMA-6. (f) Histogram showing $g_{iso}^{calc} - g_{iso}^{pred}$ for structures in TE-2.

therefore, pinpointing specific interactions in the case of H−H is challenging. Among features which encode angles formed by three atoms, H−C−C and N−C−O angles were found to be important for ERT-MBTR (see Figure S4b in the Supporting Information). The importance of H−C−C and N−C−O features is possibly related to the conformational changes in the six-membered ring of the radical moiety. This suggestion is also supported by the observation that C−N−O angles as well as C−C−C−C and C−N−C−C dihedral angles are important features in the case of ERT-DAD (see Figure S3b,c in the Supporting Information). Features in the k_1 dimension of MBTR, i.e., atomic numbers, were not found to be important in comparison to k_2 and k_3 features, possibly due to the **TR** containing molecular structures made up of the same type of atoms. However, these features might become important if a model is trained on data containing different radical moieties, differing in their type of atomic species. For both ERT models, while the most important features remained the same, their relative importance was found to change with the shuffling of **TR** and initializing the learning algorithm with different random states. Moreover, feature importance scores may be biased toward features with high cardinality or a higher number of unique values in comparison to other feature types (see Section A3 in the Supporting Information).[44,45,90] For instance, as element H has the highest number of atoms in PTMA and long-range interactions are included in MBTR, possible values of H−H distances as a feature can take multiple, unique values, leading to a larger possibility of splits using this feature and a biased importance score. Therefore, only a qualitative discussion of features with consistently high importance scores was done, and a quantitative comparison of different features based on feature importance scores was not attempted.

To summarize this section, the ERT model trained on MBTR feature vectors affords predictions of g_{iso} with mean deviations from g_{iso}^{calc} of the order of 0.0001 which is comparable to the magnitude of variation in g_{iso}^{calc} per time frame. The obtained errors are also comparable to experimentally accessible g shifts, which are usually of the order of 10^{-4}. Furthermore, the model was found to be sensitive to the radical density, as well as to the extent of equilibration.

Application to MD-2. The final ERT-MBTR model was applied to structures from MD-2 to evaluate its ability to reproduce the evolution of g_{iso} along unknown MD trajectories. MD-2 differs from MD-1 in terms of the number of oligomer molecules used in the simulation. MD-2 with each PTMA-X (X = 1, 2, 3, 4, and 6) consists of only one oligomer molecule in the simulation cell instead of 24 in the case of MD-1. Therefore, the oligomer evolves under the influence of only intramolecular interactions and its interactions with the electrolyte. Furthermore, the model is evaluated on oligomer

https://doi.org/10.1021/acs.jctc.3c01252
J. Chem. Theory Comput. 2024, 20, 2592−2604

molecules that are structurally dissimilar from structures in **TR** in terms of the number of monomer units that are radicals. For this purpose, two unknown radical densities, PTMA-2 and PTMA-4, with one less and one more radical unit, respectively, in comparison to PTMA-3 were selected. For **TE-2**, the structures are extracted only from the initialization step of each PTMA-*X* MD trajectory, as larger g deviations occur only during the initial time frames of the MD trajectory, and g_{iso}^{calc} was found to converge already during the initialization step.

Figure 5 summarizes the predictive ability of the model for structures in **TE-2**. For PTMA-1, PTMA-3, and PTMA-6, the evolution of g_{iso}^{pred} is in excellent agreement with g_{iso}^{calc} (see Figure 5a,c,e). While comparatively larger mean deviations were observed for PTMA-2 and PTMA-4 (see Figure 5b,d), the evolution of g_{iso} is fairly well reproduced and MAE/RMSE values did not exceed 0.0002 (see Table 3) for any radical

Table 3. Error Metrics Obtained for the ERT-MBTR Model in the Case of the TE-2 Data Set and Individual Radical Densities

data set	R^2	MAE [$\times 10^{-4}$]	RMSE [$\times 10^{-4}$]
TE-2	0.989	1.15	1.38
PTMA-1	0.992	0.95	1.14
PTMA-2	0.981	1.54	1.65
PTMA-3	0.993	0.95	1.02
PTMA-4	0.982	1.31	1.73
PTMA-6	0.991	1.02	1.23

density present in **TE-2**. For the whole **TE-2** data set, the order of deviation of g_{iso}^{pred} from g_{iso}^{calc} (see Figure 5f) is similar to the deviations observed for **TE-1** with both MAE and RMSE values of the order of 0.0001 (see Figure 3b,d). This indicates that the ERT-MBTR model generalizes well toward unknown MD trajectories. Additionally, for each radical density in **TE-2**, the model effectively differentiates between the structures at the beginning of the simulation and the equilibrated structures once the convergence of g_{iso}^{calc} is achieved.

Better predictive performance toward PTMA-1, PTMA-3, and PTMA-6 is expected due to the composition of the **TR** data set used for training, which consists of these radical densities. However, the ability of the model to interpolate to other radical densities that are not included in the **TR** data set is found to be sufficient for the application. The largest MAE was obtained for PTMA-2, which did not exceed 2×10^{-4} (see Table 3). In comparison, MAE values are slightly lower for PTMA-4 but RMSE indicates deviations that are larger in magnitude. In both cases, deviations seem to be more prominent after the convergence of g_{iso}^{calc} occurs, or equivalently, after the system reaches equilibrium and only minor variations in structural features occur. A comparison of MBTR outputs of structures from different time frames reveals that while structures before equilibrium are considerably different from equilibrated structures, only minor differences in the structural features are observed after g_{iso}^{calc} converges (see Figure S4 in the Supporting Information). Nevertheless, the errors are smaller than the experimental g shifts observed for PTMA and comparable to environment-dependent g shifts measurable by high-field EPR in the case of organic radicals. The ability of the model to differentiate between equilibrated structures indicates that subtle structural changes are identified by the model, and consequently, minor differences in g_{iso}^{calc} after convergence is reached are also reproduced well. Therefore, in addition to predictive accuracy with respect to absolute g_{iso} values, g shifts occurring for smaller changes in the state of charge can also be targeted. More fine-grained states of charge are still feasible through MD, and in this case, the ML approach serves as a more scalable method than DFT to predict g shifts for a larger set of structures.

CONCLUSIONS

The applicability of ML methods to predict g_{iso}, an EPR observable, from the structural features of PTMA, an organic radical polymer, was examined. PTMA molecules were dynamically evolved in an electrolyte using classical MD simulations, and oligomer structures were derived from the MD trajectory. Molecular descriptors, which encode either local or global structural features, were used to represent the molecular structure. A model based on regression trees was trained and the dependence of model performance on the encoding of structural features was studied. In the case of PTMA, global molecular descriptors such as MBTR, which encode the whole structure, were found to be more suitable. Mean deviations of ML-predicted g values (g_{iso}^{pred}) from DFT-calculated g values (g_{iso}^{calc}) were of the order of 1×10^{-4}, which is accurate enough to detect structure- and environment-dependent g shifts in the case of PTMA. Furthermore, the evolution of g_{iso}^{pred} along an unknown MD trajectory showed remarkable agreement with the evolution of g_{iso}^{calc}, thereby making the proposed ML-based approach a viable method to predict g and validate MD protocols for PTMA and similar systems. As the total computational cost of training and prediction using the ML model is lower than DFT, the approach offers a method with better scalability to predict g_{iso}. Consequently, g_{iso} shifts between transient states of the radical polymer observed experimentally using EPR and simulated using MD can be compared. Since the agreement between g_{iso}^{pred} and experimental g values is tied to the data used for training, MD simulations of the ORB system need to account for complex active material environments. As polymer conformations may be affected by interactions with other polymer chains, g value variations may also be influenced by interchain effects. A better representation of the experimental system may be achieved by the inclusion of multiple polymer chains and other electrode constituents such as conductive additives. In this aspect, the presented protocol can be utilized to compare MD protocols and test for agreement with an experimental system by using g as an experimental observable. A further refinement of the simulated system based on insights from EPR experiments can improve the agreement between g_{iso}^{calc} and experimental g values also for low radical densities, which, in turn, should afford g_{iso}^{pred} consistent with experimental g values. As the variation in g_{iso} with the radical density is larger than the error of the ML model, experimental g values can be used to benchmark such a combination of MD and ML. The protocol used in this work can be extended to larger simulation systems by conducting predictions on smaller subsystems and obtaining a distribution of g values for further statistical analysis and validation of large-scale MD simulations. Furthermore, the workflow can be transferred to other paramagnetic species provided that the g values can be experimentally verified and theoretically computed with sufficient accuracy. While isotropic parameters and their distributions can be targeted using the current protocol, transferability to tensorial properties may need to be investigated separately.

https://doi.org/10.1021/acs.jctc.3c01252
J. Chem. Theory Comput. 2024, 20, 2592−2604

The ML workflow utilized in this work prioritizes low computational costs and uses a limited number of structures for training. While further gains in prediction accuracy may be achieved by increasing the size of the **TR**, the obtained accuracy is in line with the aims of the work. Upon application to unknown structures, the predictive accuracy was found to be higher for radical densities included in the **TR**s. Therefore, performance gains may be achieved by including additional radical densities in the **TR**. In order to extend the scope of the model, data-efficient protocols must be developed. Through an optimization of the **TR**, the approach can be extended to MD simulations which may vary from the simulations shown in the current work, in terms of simulation parameters such as temperature, chain length of the polymer chain, or radical moieties. To realize such functional extensibility, the ML workflow demonstrated in this work can be adapted into an active-learning ML workflow in which structures sourced from MD trajectories are continuously screened with respect to their dissimilarity from the current **TR** so that dissimilar structures can be added to a modified **TR** and used to retrain the model. Notably, we envision that the workflow can be adapted to other redox-active organic materials and that other properties relevant to organic radical polymer batteries, such as electron coupling parameters, may be targeted. As the predicted g values remain sensitive to the radical density and structural changes, the state of charge-dependent g shifts for ORB setups may become computationally accessible and may be applied to even larger simulation systems.

■ ASSOCIATED CONTENT

Data Availability Statement

Data sets, trained models, and scripts used in this work are available at Jülich DATA (https://doi.org/10.26165/JUELICH-DATA/TOBXWP) and Jülich Gitlab (https://jugit.fz-juelich.de/iek_9_spectroscopy/ptma-ml).

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.3c01252.

> Description of cross-validation, cross-validated error metrics for various regressors, optimized hyperparameters, description of error metrics, feature importance scores of the trained ERT-MBTR model, ESP fits, and details of *in operando* EPR experiments (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Davis Thomas Daniel − *Institute of Energy and Climate Research (IEK-9), Forschungszentrum Jülich GmbH, 52425 Jülich, Germany; Institute of Technical and Macromolecular Chemistry, RWTH Aachen University, 52056 Aachen, Germany;* orcid.org/0000-0001-7035-3416; Email: d.daniel@fz-juelich.de

Authors

Souvik Mitra − *Institute of Physical Chemistry, University of Münster, 48149 Münster, Germany*

Rüdiger-A. Eichel − *Institute of Energy and Climate Research (IEK-9), Forschungszentrum Jülich GmbH, 52425 Jülich, Germany; Institute of Physical Chemistry, RWTH Aachen University, Aachen 52056, Germany;* orcid.org/0000-0002-0013-6325

Diddo Diddens − *Helmholtz Institute Münster (IEK-12), Forschungszentrum Jülich GmbH, 48149 Münster, Germany;* orcid.org/0000-0002-2137-1332

Josef Granwehr − *Institute of Energy and Climate Research (IEK-9), Forschungszentrum Jülich GmbH, 52425 Jülich, Germany; Institute of Technical and Macromolecular Chemistry, RWTH Aachen University, 52056 Aachen, Germany*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.3c01252

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Dr. Christoph Scheurer for valuable comments regarding the experimental approach and Dr. Conrad Szczuka for helpful discussion. The authors also gratefully acknowledge Steffen Oevermann for preparation of PTMA cathode films used for EPR experiments, Mary Joan Blümich for editing the abstract and Dr. Simone Köcher for her guidance with research data management and curation. Computational resources from the RWTH Aachen University under project rwth1253 are acknowledged. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within the priority program "Polymer-Based Batteries" (SPP 2248)—project no. 441255373.

■ REFERENCES

(1) Goswami, M.; Chirila, A.; Rebreyend, C.; de Bruin, B. EPR spectroscopy as a tool in homogeneous catalysis research. *Top. Catal.* **2015**, *58*, 719−750.

(2) Neese, F. *eMagRes*; John Wiley & Sons, Ltd, 2017, pp 1−22.

(3) Neese, F. Quantum chemical calculations of spectroscopic properties of metalloproteins and model compounds: EPR and Mössbauer properties. *Curr. Opin. Chem. Biol.* **2003**, *7*, 125−135.

(4) Patchkovskii, S.; Ziegler, T. Calculation of the EPR g-tensors of high-spin radicals with density functional theory. *J. Phys. Chem. A* **2001**, *105*, 5490−5497.

(5) Stoll, S. *Multifrequency Electron Paramagnetic Resonance*, Chapter 3; John Wiley & Sons, Ltd, 2014, pp 69−138.

(6) Schmidt, J.; Marques, M. R.; Botti, S.; Marques, M. A. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **2019**, *5*, 83.

(7) Keith, J. A.; Vassilev-Galindo, V.; Cheng, B.; Chmiela, S.; Gastegger, M.; Müller, K. R.; Tkatchenko, A. Combining machine learning and computational chemistry for predictive insights into chemical systems. *Chem. Rev.* **2021**, *121*, 9816−9872.

(8) Meuwly, M. Machine learning for chemical reactions. *Chem. Rev.* **2021**, *121*, 10218−10239.

(9) Paruzzo, F. M.; Hofstetter, A.; Musil, F.; De, S.; Ceriotti, M.; Emsley, L. Chemical shifts in molecular solids by machine learning. *Nat. Commun.* **2018**, *9*, 4501.

(10) Himmetoglu, B. Tree based machine learning framework for predicting ground state energies of molecules. *J. Chem. Phys.* **2016**, *145*, 134101.

(11) Okamoto, Y.; Kubo, Y. Ab initio calculations of the redox potentials of additives for lithium-ion batteries and their prediction through machine learning. *ACS Omega* **2018**, *3*, 7868−7874.

(12) Ermler, W. C.; Ross, R. B.; Christiansen, P. A. *Advances in Quantum Chemistry*; Elsevier, 1988; Vol. *19*, pp 139−182.

(13) Weil, J.; Bolton, J., Chapter 4. *Electron Paramagnetic Resonance*; Weil, J., Bolton, J., Eds.; John Wiley & Sons, Ltd, 2006; pp 85−117.

(14) Gast, P.; Groenen, E. J. *eMagRes*; John Wiley & Sons, Ltd, 2016; pp 1435−1444.
(15) Solanki, N. K.; McInnes, E. J.; Mabbs, F. E.; Radojevic, S.; McPartlin, M.; Feeder, N.; Davies, J. E.; Halcrow, M. A. Steric Control of the Electronic Ground State in Six-Coordinate Copper (ii) Complexes. *Angew. Chem., Int. Ed.* **1998**, *37*, 2221−2223.
(16) Miner, K. D.; Pfister, T. D.; Hosseinzadeh, P.; Karaduman, N.; Donald, L. J.; Loewen, P. C.; Lu, Y.; Ivancich, A. Identifying the elusive sites of tyrosyl radicals in cytochrome c peroxidase: implications for oxidation of substrates bound at a site remote from the heme. *Biochemistry* **2014**, *53*, 3781−3789.
(17) Stoll, S.; Shafaat, H. S.; Krzystek, J.; Ozarowski, A.; Tauber, M. J.; Kim, J. E.; Britt, R. D. Hydrogen bonding of tryptophan radicals revealed by EPR at 700 GHz. *J. Am. Chem. Soc.* **2011**, *133*, 18098−18101.
(18) Un, S.; Atta, M.; Fontecave, M.; Rutherford, A. W. g-Values as a probe of the local protein environment: high-field EPR of tyrosyl radicals in ribonucleotide reductase and photosystem II. *J. Am. Chem. Soc.* **1995**, *117*, 10713−10719.
(19) Ivancich, A.; Mattioli, T. A.; Un, S. Effect of protein microenvironment on tyrosyl radicals. A high-field (285 GHz) EPR, resonance Raman, and hybrid density functional study. *J. Am. Chem. Soc.* **1999**, *121*, 5743−5753.
(20) Anderson, P.; Weiss, P. Exchange Narrowing in Paramagnetic Resonance. *Rev. Mod. Phys.* **1953**, *25*, 269−276.
(21) Oyaizu, K.; Nishide, H. Radical polymers for organic electronic devices: a radical departure from conjugated polymers? *Adv. Mater.* **2009**, *21*, 2339−2344.
(22) Janoschka, T.; Hager, M. D.; Schubert, U. S. Powering up the future: radical polymers for battery applications. *Adv. Mater.* **2012**, *24*, 6397−6409.
(23) Suga, T.; Ohshiro, H.; Sugita, S.; Oyaizu, K.; Nishide, H. Emerging N-Type Redox-Active Radical Polymer for a Totally Organic Polymer-Based Rechargeable Battery. *Adv. Mater.* **2009**, *21*, 1627−1630.
(24) Wild, A.; Strumpf, M.; Häupler, B.; Hager, M. D.; Schubert, U. S. All-organic battery composed of thianthrene- and TCAQ-based polymers. *Adv. Energy Mater.* **2017**, *7*, 1601415.
(25) Muench, S.; Burges, R.; Lex-Balducci, A.; Brendel, J. C.; Jäger, M.; Friebe, C.; Wild, A.; Schubert, U. S. Printable ionic liquid-based gel polymer electrolytes for solid state all-organic batteries. *Energy Storage Mater.* **2020**, *25*, 750−755.
(26) Neese, F. Prediction of electron paramagnetic resonance g values using coupled perturbed Hartree−Fock and Kohn−Sham theory. *J. Chem. Phys.* **2001**, *115*, 11080−11096.
(27) Schreckenbach, G.; Ziegler, T. Density functional calculations of NMR chemical shifts and ESR g-tensors. *Theor. Chem. Acc.* **1998**, *99*, 71−82.
(28) Schreckenbach, G.; Ziegler, T. Calculation of the g-tensor of electron paramagnetic resonance spectroscopy using gauge-including atomic orbitals and density functional theory. *J. Phys. Chem. A* **1997**, *101*, 3388−3399.
(29) Malkina, O. L.; Vaara, J.; Schimmelpfennig, B.; Munzarová, M.; Malkin, V. G.; Kaupp, M. Density Functional Calculations of Electronic g-Tensors Using Spin−Orbit Pseudopotentials and Mean-Field All-Electron Spin- Orbit Operators. *J. Am. Chem. Soc.* **2000**, *122*, 9206−9218.
(30) Angstl, R. Contribution of the relativistic mass correction to the g tensor of molecules. *Chem. Phys.* **1989**, *132*, 435−442.
(31) Stone, A. J. Gauge invariance of the g tensor. *Proc. R. Soc. London, Ser. A* **1963**, *271*, 424−434.
(32) Harriman, J. E. *Theoretical Foundations of Electron Spin Resonance*; Academic Press, 2013.
(33) Neese, F. Software update: The ORCA program system—Version 5.0. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2022**, *12*, No. e1606.
(34) Daniel, D. T.; Oevermann, S.; Mitra, S.; Rudolf, K.; Heuer, A.; Eichel, R.-A.; Winter, M.; Diddens, D.; Brunklaus, G.; Granwehr, J. Multimodal investigation of electronic transport in PTMA and its impact on organic radical battery performance. *Sci. Rep.* **2023**, *13*, 10934.
(35) Bowler, D. R.; Miyazaki, T. \mathcal{O}(N) methods in electronic structure calculations. *Rep. Prog. Phys.* **2012**, *75*, 036503.
(36) Goedecker, S. Linear scaling electronic structure methods. *Rev. Mod. Phys.* **1999**, *71*, 1085−1123.
(37) Nakata, A.; Baker, J. S.; Mujahed, S. Y.; Poulton, J. T.; Arapan, S.; Lin, J.; Raza, Z.; Yadav, S.; Truflandier, L.; Miyazaki, T.; et al. Large scale and linear scaling DFT with the CONQUEST code. *J. Chem. Phys.* **2020**, *152*, 164112.
(38) Mohr, S.; Ratcliff, L. E.; Genovese, L.; Caliste, D.; Boulanger, P.; Goedecker, S.; Deutsch, T. Accurate and efficient linear scaling DFT calculations with universal applicability. *Phys. Chem. Chem. Phys.* **2015**, *17*, 31360−31370.
(39) Niemöller, A.; Jakes, P.; Eurich, S.; Paulus, A.; Kungl, H.; Eichel, R.-A.; Granwehr, J. Monitoring local redox processes in $LiNi_{0.5}Mn_{1.5}O_4$ battery cathode material by in operando EPR spectroscopy. *J. Chem. Phys.* **2018**, *148*, 014705.
(40) Kulikov, I.; Panjwani, N. A.; Vereshchagin, A. A.; Spallek, D.; Lukianov, D. A.; Alekseeva, E. V.; Levin, O. V.; Behrends, J. Spins at work: probing charging and discharging of organic radical batteries by electron paramagnetic resonance spectroscopy. *Energy Environ. Sci.* **2022**, *15*, 3275−3290.
(41) Kanzaki, Y.; Mitani, S.; Shiomi, D.; Morita, Y.; Takui, T.; Sato, K. Microscopic Behavior of Active Materials Inside a TCNQ-Based Lithium-Ion Rechargeable Battery by in Situ 2D ESR Measurements. *ACS Appl. Mater. Interfaces* **2018**, *10*, 43631−43640.
(42) Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Müller, K. R. Bypassing the Kohn-Sham equations with machine learning. *Nat. Commun.* **2017**, *8*, 872.
(43) Margraf, J. T.; Reuter, K. Making the coupled cluster correlation energy machine-learnable. *J. Phys. Chem. A* **2018**, *122*, 6343−6348.
(44) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5−32.
(45) Breiman, L.; Friedman, J.; Stone, C. J.; Olshen, R. A. *Classification and Regression Trees*; CRC Press, 1984.
(46) Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3−42.
(47) Chen, A. Y.; Lee, J.; Damjanovic, A.; Brooks, B. R. Protein p K a Prediction by Tree-Based Machine Learning. *J. Chem. Theory Comput.* **2022**, *18*, 2673−2686.
(48) Deringer, V. L.; Bartók, A. P.; Bernstein, N.; Wilkins, D. M.; Ceriotti, M.; Csányi, G. Gaussian process regression for materials and molecules. *Chem. Rev.* **2021**, *121*, 10073−10141.
(49) Bahlke, M. P.; Mogos, N.; Proppe, J.; Herrmann, C. Exchange spin coupling from Gaussian process regression. *J. Phys. Chem. A* **2020**, *124*, 8708−8723.
(50) Meiler, J. PROSHIFT: protein chemical shift prediction using artificial neural networks. *J. Biomol. NMR* **2003**, *26*, 25−37.
(51) Taguchi, A. T.; Evans, E. D.; Dikanov, S. A.; Griffin, R. G. Convolutional Neural Network Analysis of Two-Dimensional Hyperfine Sublevel Correlation Electron Paramagnetic Resonance Spectra. *J. Phys. Chem. Lett.* **2019**, *10*, 1115−1119.
(52) Worswick, S. G.; Spencer, J. A.; Jeschke, G.; Kuprov, I. Deep neural network processing of DEER data. *Sci. Adv.* **2018**, *4*, No. eaat5218.
(53) Szczuka, C.; Eichel, R.-A.; Granwehr, J. Gauging the importance of structural parameters for hyperfine coupling constants in organic radicals. *RSC Adv.* **2023**, *13*, 14565−14574.
(54) Langer, M. F.; Goeßmann, A.; Rupp, M. Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning. *npj Comput. Mater.* **2022**, *8*, 41.
(55) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115.
(56) Himanen, L.; Jäger, M. O.; Morooka, E. V.; Federici Canova, F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **2020**, *247*, 106949.

https://doi.org/10.1021/acs.jctc.3c01252
J. Chem. Theory Comput. 2024, 20, 2592−2604

(57) Huo, H.; Rupp, M. Unified representation of molecules and crystals for machine learning. *Mach. Learn.: Sci. Technol.* **2022**, *3*, 045017.
(58) Cheng, L.; Sun, J.; Miller, T. F. Accurate molecular-orbital-based machine learning energies via unsupervised clustering of chemical space. *J. Chem. Theory Comput.* **2022**, *18*, 4826−4835.
(59) Klem, H.; Hocky, G. M.; McCullagh, M. Size-and-shape space gaussian mixture models for structural clustering of molecular dynamics trajectories. *J. Chem. Theory Comput.* **2022**, *18*, 3218−3230.
(60) Scikit-learn. Gaussian mixture model. https://scikit-learn.org/stable/modules/mixture.html# (accessed January 10, 2024).
(61) McLachlan, G. J.; Lee, S. X.; Rathnayake, S. I. Finite mixture models. *Annu. Rev. Stat. Appl.* **2019**, *6*, 355−378.
(62) Meng, X.-L.; Van Dyk, D. The EM algorithm—an old folk-song sung to a fast new tune. *J. R. Statist. Soc. B* **1997**, *59*, 511−567.
(63) Brown, J. D.; Summers, M. F.; Johnson, B. A. Prediction of hydrogen and carbon chemical shifts from RNA using database mining and support vector regression. *J. Biomol. NMR* **2015**, *63*, 39−52.
(64) Kuhn, S.; Egert, B.; Neumann, S.; Steinbeck, C. Building blocks for automated elucidation of metabolites: Machine learning methods for NMR prediction. *BMC Bioinf.* **2008**, *9*, 400.
(65) Rapaport, D. *The Art of Molecular Dynamics Simulation*; Cambridge University Press, 2004.
(66) Canongia Lopes, J. N.; Deschamps, J.; Pádua, A. A. H. Modeling ionic liquids using a systematic all-atom force field. *J. Phys. Chem. B* **2004**, *108*, 2038−2047.
(67) Canongia Lopes, J. N.; Padua, A. A. H. Molecular force field for ionic liquids III: imidazolium, pyridinium, and phosphonium cations; chloride, bromide, and dicyanamide anions. *J. Phys. Chem. B* **2006**, *110*, 19586−19592.
(68) Perner, V.; Diddens, D.; Otteny, F.; Küpers, V.; Bieker, P.; Esser, B.; Winter, M.; Kolek, M. Insights into the solubility of poly (vinylphenothiazine) in carbonate-based battery electrolytes. *ACS Appl. Mater. Interfaces* **2021**, *13*, 12442−12453.
(69) Mitra, S.; Heuer, A.; Diddens, D. Electron transfer reaction of TEMPO-based organic radical batteries in different solvent environments: comparing quantum and classical approaches. *Phys. Chem. Chem. Phys.* **2024**, *26*, 3020−3028.
(70) Martínez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M. PACKMOL: A package for building initial configurations for molecular dynamics simulations. *J. Comput. Chem.* **2009**, *30*, 2157−2164.
(71) Abraham, M.; Alekseenko, A.; Bergh, C.; Blau, C.; Briand, E.; Doijade, M.; Fleischmann, S.; et al. *GROMACS 2023 Manual*, 2023.
(72) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.
(73) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; et al. *Gaussian 16*, Revision C.01; Gaussian Inc.: Wallingford, CT, 2016.
(74) Atkins, P.; Friedman, R. *Molecular Quantum Mechanics*; OUP Oxford, 2011.
(75) Dunning, T. H., Jr. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007−1023.
(76) Neese, F. Software update: the ORCA program system, version 4.0. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2018**, *8*, No. e1327.
(77) Adamo, C.; Cossi, M.; Barone, V. An accurate density functional method for the study of magnetic properties: the PBE0 model. *J. Mol. Struct.: THEOCHEM* **1999**, *493*, 145−157.
(78) Stoychev, G. L.; Auer, A. A.; Neese, F. Automatic generation of auxiliary basis sets. *J. Chem. Theory Comput.* **2017**, *13*, 554−562.
(79) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(80) Hjorth Larsen, A.; Jørgen Mortensen, J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C.; et al. The atomic simulation environment—a Python library for working with atoms. *J. Phys.: Condens. Matter* **2017**, *29*, 273002.
(81) Li, H.; Liang, Y.; Xu, Q. Support vector machines and its applications in chemistry. *Chemom. Intell. Lab. Syst.* **2009**, *95*, 188−198.
(82) Geurts, P.; Irrthum, A.; Wehenkel, L. Supervised learning with decision tree-based methods in computational and systems biology. *Mol. BioSyst.* **2009**, *5*, 1593−1605.
(83) Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123−140.
(84) Freund, Y.; Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119−139.
(85) Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367−378.
(86) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.
(87) Daniel, D. T.; Szczuka, C.; Jakes, P.; Eichel, R.-A.; Granwehr, J. Laplace inverted pulsed EPR relaxation to study contact between active material and carbon black in Li−organic battery cathodes. *Phys. Chem. Chem. Phys.* **2023**, *25*, 12767−12776.
(88) Kemper, T. W.; Larsen, R. E.; Gennett, T. Relationship between molecular structure and electron transfer in a polymeric nitroxyl-radical energy storage material. *J. Phys. Chem. C* **2014**, *118*, 17213−17220.
(89) Bobela, D. C.; Hughes, B. K.; Braunecker, W. A.; Kemper, T. W.; Larsen, R. E.; Gennett, T. Close packing of nitroxide radicals in stable organic radical polymeric materials. *J. Phys. Chem. Lett.* **2015**, *6*, 1414−1419.
(90) Strobl, C.; Boulesteix, A.-L.; Zeileis, A.; Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinf.* **2007**, *8*, 25.

https://doi.org/10.1021/acs.jctc.3c01252
J. Chem. Theory Comput. 2024, 20, 2592−2604